Report of the CSAC Differential Privacy Working Group- Tasks 1, 2 & 3

Richelle Winkler (convenor), Jay Breidt (ex officio), John Czajka, Barbara Entwisle, Kunal Talwar, Lance Waller, Claire McKay Bowen, Ron Prevost

Charter for CSAC Differential Privacy Working Group

Four tasks, with deliverables consisting of presentations at CSAC

Focus on DHC, Detailed DHC-A, Detailed DHC-B, S-DHC

- **Task 1. Developing a summary of use cases.**
- **Task 2. Developing recommendations for prioritizing use cases for the administration of a "privacy-loss budget."**
- **Task 3. Developing metrics to assess the impact of differential privacy on the accuracy of decennial census data.**
- **Task 4. Developing strategies for communicating the use of differential privacy for the 2020 Census data products.**

Request for Feedback

Specific comments/suggestions on the DHC demonstration data released 8/25/2022

Overall Comment from Working Group

DRAFT RECOMMENDATION: For Census 2030 and all other Census data products, the Census Bureau should have a clear, well-tested, practical, detailed and step-by-step plan for how the DAS system will be applied for each data product and how such application at each level will impact other products/levels prior to releasing any data.

Task 1- Use Case Identification- Activities

- Met with FSCPE and reviewed materials they collected on use cases
- Participated in discussions and workshops organized by Census Quality Reinforcement (CQR) Task Force
- Participated in CNSTAT workshop on DHC demo in June 2022
- Reviewed feedback collected by Census Bureau on DHC demonstration data (March 2022 release)
- Reviewed published materials and presentations about data uses and DP implications

Task 1- Reviewing Cases- Findings

- **Census is a general use tool, thousands (millions?) of use cases**
  - Probably thousands of funding allocation decisions for federal, state, and local decisions
  - Census is only good source of demographic data for rural areas
  - Valuable *because of high accuracy for small geographies*- distinguishes decennial census
- **Use cases collected are limited to those with resources, expertise, time, awareness, and patience to contribute.**
- **Burn-out, file layout, mistakes, and deadlines limit responses**

Task 1- Reviewing Cases- Findings

- **Block accuracy does matter for basic pop counts and by race/ethnicity**
  - Blocks allow for custom geography creation
  - Census is only source
  - Redistricting, transportation planning, emergency response, environmental justice
  - Blocks build (or lose) trust among users, because people can ground truth

DRAFT RECOMMENDATION: Planning for Census 2030, consider changing construction of census blocks so that none are so small that accurate population totals can't be released OR use another DAS method for certain variables, like population totals.

Task 1- Use Cases- Findings

- **Application of DP throughout Census 2020, and still in the most recent DHC demo data (8/25/22), significantly impacts the accuracy of data for small domains:**
  - Rural areas
  - Groups other than NH white
  - Small populations (ex. same-sex couples, small municipalities, etc.)
    - Limited accuracy below state level

Draft Recommendations on Process

- *The Census Bureau should document and communicate the impact of DP application on these groups (rural, BIPOC, small pops), and use the results to improve Census 2030 and other data products.*
- *The Census Bureau should develop a research program that seeks to (a) improve privacy protection methodologies based on the results from the impacts of DP applications and (b) link persons to households.*
- *This research program should include collaboration with user groups at the local and state levels to define the utility of DHC variables for safely accessing the data at various geographic levels early in the decennial census development cycle (ex. adaptive methods that provide different detail for different geographies or pop sizes).*

Task 2- Prioritizing Use Cases

- **Age group (5-year) accuracy is particularly important**
  - County age-specific health rates (CDC and others)
  - Basic county-level governmental planning for services
  - Making accurate population estimates and projections

DRAFT RECOMMENDATIONS:

- ***Allocate greater accuracy (PLB) to county x age so accurate even for more rural counties in 2020 DHC, and to place x age to improve accuracy for places.***
- ***Consider collapsing older age groups (85-99, 100+).***
- ***Plan for Census 2030, allocate more PLB to age (under/over 18) in the Redistricting data.***

Task 2- Prioritizing Use Cases

- **Tenure by race/ethnicity/age -- Fair Housing Act, Lending**
  - **Race.** Except for White alone, MAPE ranges from 47% to 238% for owner-occupied and 46% to 167% for renter-occupied units in counties (Table 4.b). Similar for places.
  - **Ethnicity.** (Table 3.b/c). Hispanic HH tenure in incorporated places, MAPEs= 44% for owner- and renter-occupied. More than half the units have absolute percentage differences >5%.
  - **Age.** Tenure by age group of householder for incorporated places (Table 2.c). For renter occupied housing units, MAPE= 20%-47%, and well over half of housing units in each age group have absolute percentage differences >5%.

DRAFT RECOMMENDATION: Allocate greater accuracy (PLB) to tenure by race/ethnicity and age group for counties, places, and tracts.

Task 2- Prioritizing Use Cases

- **Housing Units Reason for Vacancy**
  - We have a national housing crisis that is playing out differently and more acutely in some local areas more than others. We need to know why vacant at local level. Different reasons (For sale/rent, migrant worker, seasonal) mean very different implications.
    - Ex.) Growth of short-term rental market. Units are vacant, but they might be high value seasonal homes or they might be available for rent. Need to know difference to make decisions about new construction, zoning, planning, how to regulate short-term rentals. Important for labor markets.
  - For all incorporated places, MAPE ranges from 16 to 52 percent depending on the reason (Table 5.c)

Task 2- Prioritizing Use Cases

- **Housing Units Reason for Vacancy (cont.)**
  - Seasonal housing data used ***at block level*** for fire planning (US Forest Service) and by lake associations for planning, for school district funding planning, and local govt zoning/planning.
  - Local demographers need these data on vacancy rates to estimate seasonal rates. Consider seasonal/not.

DRAFT RECOMMENDATION: Allocate greater accuracy (PLB) to reason for vacancy, at least to distinguish seasonal/recreational, migrant, and other types.

Task 2- Prioritizing Use Cases

- **Accurate persons per HH for counties, places, MCD, and tracts**
  - In order to make population estimates and projections using housing unit method
  - Variable not direct in DHC but users may calculate and find impossible/improbable results

DRAFT RECOMMENDATIONS:

- ***For Census 2020, direct users to the Detailed DHC or S-DHC for this purpose and focus Detailed DHC/S-DHC efforts to prioritize this variable.***
- ***For 2030, the Census Bureau should set a research agenda for how to better match persons-households to improve this accuracy.***

Task 2- Prioritizing Use Cases

- **Less priority to number of races**
  - No clearly identified use cases for Tables 10a/10b/10c/10d and 11a/b/c/d
  - High frequency of absolute percentage difference exceeding 10% at county, incorporated places and tract levels

DRAFT RECOMMENDATION: Remove number of races from DHC and keep the number of races in Detailed DHC.

Task 3- Metrics to Assess DP Impact

- Metrics are critical. Many users don’t have capacity to review the raw data.
- It remains important to identify outliers where errors may be more extreme.
- Metrics may be misinterpreted, as data users tend to treat the swapped data as the “truth”.
- PLB is still not intuitive for data users and under researched on what it means for various data types and applications in literature.
- The CSAC DP WG has made several recommendations in the past. We have few remaining.

Draft Recommendations on Metrics

- *For cases where there are many units with an error of 5% or more, add a column to also show the number of units with an error of 10% or more.*
- *Reevaluate the privacy-loss budget changes using a traditional measure of disclosure risk to help provide context of what a privacy loss budget means.*
- *Create an independent group (including users) that can access the CEF to evaluate the privacy-utility trade-off. If this is not feasible, the Census Bureau should provide some illustrative examples, perhaps using qualitative measures, of how error metrics based on the CEF would compare.*

Discussion